Computer experiment 3 (Ch5, Ch6, [Theodoridis 2009])

1.

a. (i) Generate four sets, each one consisting of 100 two-dimensional vectors, from the normal distributions with mean values $[-10,\ -10]^T$, $[-10,\ 10]^T$, $[10,\ -10]^T$, $[10,\ 10]^T$ and covariance matrices equal to $0.2 * I$. These sets constitute the data set for a four-class two-dimensional classification problem (each set corresponds to a class).

a. (ii) Compute the Sw, Sb, and Sm scatter matrices.

a. (iii) Compute the value for the criterion $J3$.

b. Repeat (a) when the mean vectors of the normal distributions that generate the data are $[-1,\ -1]^T$, $[-1,\ 1]^T$, $[1,\ -1]^T$, $[1,\ 1]^T$.

c. Repeat (a) when the covariance matrices of the normal distributions that generate the data are equal to $3 * I$.

2. The Fisher's discriminant ratio (FDR) is defined by:

$$FDR = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}.$$

FDR is sometimes used to quantify the separability capabilities of **individual** features.

a. (i) Generate two sets, each one consisting of 100 two-dimensional vectors, from the normal distributions with mean values $[2,\ 4]^T$ and $[2.5,\ 10]^T$ and covariance matrices equal to the 2×2 identity matrix I. Their composition forms the data set for a two class two dimensional classification problem (each set corresponds to a class).

a. (ii) Compute the value of the FDR index for both features.

b. Repeat (a) when the covariance matrices of the normal distributions that generate the data are both equal to $0.25 * I$.

c. Discuss the results.

3.

a. Generate an $l \times N$ dimensional matrix X ($l = 2$ and $N = 1000$), whose columns are two-dimensional points lying around the line h: $x_1 + x_2 = 0$

(i.e., $w = [1, 1]^T$ and $w_0 = 0$), using the *generate_hyper* function with parameters $a = 10, e = 1$ and $sed = 0$.

b. Compute the principal components of the covariance of X as well as the corresponding variances (eigenvalues). Compare the direction of the first principal component with the direction vector of h (which is perpendicular to w) and draw your conclusions.

```
function X=generate_hyper(w,w0,a,e,N,sed)
  l=length(w);

   t=(rand(l-1,N)-.5)*2*a;
   t_last=-(w(1:l-1)/w(l))'*t + 2*e*(rand(1,N)-.5)-(w0/w(l));
   X=[t; t_last];
   %Plots for the 2d and 3d case
   if(l==2)
     figure(1), plot(X(1,:),X(2,:),'.b')
   elseif(l==3)
     figure(1), plot3(X(1,:),X(2,:),X(3,:),'.b')
   end
   figure(1), axis equal
```

MATLAB function named *generate_hyper* that generates randomly l-dimensional points $\boldsymbol{x}_i = [x_1(i), x_2(i), \ldots, x_l(i)]^T$ around an $(l-1)$-dimensional hyperplane H: $\boldsymbol{w}^T\boldsymbol{x} + w_0 = 0$, where $\boldsymbol{w} = [w_1, w_2, \ldots, w_l]^T$. More specifically, the function takes as inputs: (a) the parameter (column) vector $\boldsymbol{w}$ for H ($w_l \neq 0$), (b) the offset w_0 for H, (c) a positive parameter a that defines the range $[-a,\ a]$, where each one of the first $(l-1)$ coordinates of the points is uniformly distributed, (d) the positive parameter e that defines the range $[-e,\ e]$ of a uniformly distributed noise source, which is added to the term $(-w_0 - \sum_{i=0}^{l-1} w_i x_i)/w_l$ to produce the lth coordinate, (e) the number N of points to be generated, and (f) the seed *sed* for the *rand* MATLAB function. It returns an $l \times N$ dimensional matrix, X, whose columns contain the generated data points. In addition, the function plots the data points for $l = 2, 3$.